Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 3618–3622, June 1983
Biochemistry

Human adult T-cell leukemia virus: Complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA

(human leukemia virus/provirus structure/translation frames/polyadenylylation model)

MOTOHARU SEIKI, SEISUKE HATTORI, YOKO HIRAYAMA, AND MITSUAKI YOSHIDA

Department of Viral Oncology, Cancer Institute, Kami-Ikebukuro, Toshima-ku, Tokyo, Japan

Communicated by Takashi Sugimura, March 14, 1983

ABSTRACT **Human retrovirus adult T-cell leukemia virus (ATLV) has been shown to be closely associated with human adult T-cell leukemia (ATL) [Yoshida, M., Miyoshi, I. & Hinuma, Y. (1982) *Proc. Natl. Acad. Sci. USA* 79, 2031–2035]. The provirus of ATLV integrated in DNA of leukemia T cells from a patient with ATL was molecularly cloned and the complete nucleotide sequence of 9,032 bases of the proviral genome was determined. The provirus DNA contains two long terminal repeats (LTRs) consisting of 755 bases, one at each end, which are flanked by a 6-base direct repeat of the cellular DNA sequence. The nucleotides in the LTR could be arranged into a unique secondary structure, which could explain transcriptional termination within the 3′ LTR but not in the 5′ LTR. The nucleotide sequence of the provirus contains three large open reading frames, which are capable of coding for proteins of 48,000, 99,000, and 54,000 daltons. The three open frames are in this order from the 5′ end of the viral genome and the predicted 48,000-dalton polypeptide is a precursor of gag proteins, because it has an identical amino acid sequence to that of the NH_2 terminus of human T-cell leukemia virus (HTLV) p24. The open frames coding for 99,000- and 54,000-dalton polypeptides are thought to be the *pol* and *env* genes, respectively. On the 3′ side of these three open frames, the ATLV sequence has four smaller open frames in various phases; these frames may code for 10,000-, 11,000-, 12,000-, and 27,000-dalton polypeptides. Although one or some of these open frames could be the transforming gene of this virus, in preliminary analysis, DNA of this region has no homology with the normal human genome.**

Recently, retroviruses were independently isolated from human T-cell leukemias by two groups. One retrovirus is human T-cell leukemia virus (HTLV) isolated by Gallo and colleagues from patients with cutaneous T-cell lymphoma (1, 2), and the other is adult T-cell leukemia virus (ATLV) isolated from patients with adult T-cell leukemia (ATL) (3, 4). Recently, these two viruses have been shown to be closely related (5). ATLV was shown to be associated with ATL, which is a unique disease with T-cell malignancy (6), and the provirus genome was always detected in the chromosomal DNA of the leukemia cells (4). Recently, we reported molecular cloning of provirus DNA integrated in the cell line MT-1 and the nucleotide sequence of the long terminal repeat (LTR) with 754 bases (7), and we also proposed that ATLV might be distinct from other known animal retroviruses (7). From these previous observations, identification of genetic structure and the gene products seemed to be of great importance in understanding the origin of the virus and the mechanisms of leukemogenesis by this virus. For this purpose, we isolated a clone (λATK-1) of the provirus genome integrated in ATL cell DNA.

This paper reports the complete 9,032-nucleotide sequence of the proviral genome cloned in λATK-1 and the amino acid sequence predicted for the putative proteins.

MATERIALS AND METHODS

Cloning and Sequence Analysis of Provirus DNA of ATLV Integrated in Leukemia Cells. DNA was extracted from peripheral blood cells of a patient (K.K.) with ATL, digested with *Eco*RI, and separated by electrophoresis in agarose gel. DNA fractions of the 17-kilobase fragment containing the provirus were extracted, ligated to the *Eco*RI site of Charon 4A phage DNA, and subjected to *in vitro* packaging as described by Blattner *et al.* (8). Screening with viral [^{32}P]cDNA, recombinant phage λATK-1 was isolated. The DNA fragment cloned in λATK-1 was excised by *Eco*RI and cleaved into several fragments with restriction endonucleases for subcloning in plasmid pBR322. The nucleotide sequence of the fragments was determined by the procedure of Maxam and Gilbert (9).

RESULTS

Molecular Cloning and Sequence Analysis Strategy. Previously we reported the molecular cloning (λATM-1) of the provirus genome from cell line MT-1 and identified the LTR structure (7). However, this time we have isolated a new provirus clone λATK-1 directly from DNA of leukemia cells of an ATL patient for further analysis.

A simple restriction cleavage map of the inserted fragment in λATK-1 was constructed to subclone the regions containing provirus into pBR322. As shown in Fig. 1, *Bam*HI divided the viral sequence into three fragments and these were subcloned into pBR322; thus, pATK-03, pATK-06, and pATK-08 were obtained. Plasmid pATK-100, constructed from the *Pst* I fragment of the λATK-1 insert, contained two *Bam*HI junctions between the subclones described above. The plasmids pATK-03, pATK-06, and pATK-08 were digested with *Pst* I, *Sal* I, and *Sma* I, respectively, and the fragments were subjected to sequence analysis in both strands after further digestions with *Hpa* II, *Sau*3AI, *Hin*fI, or other restriction endonucleases. The determined sequences of pATK-03, pATK-06, and pATK-08 were overlapped by sequence analysis across the two *Bam*HI sites in the clone pATK-100. Fig. 2 shows the 9,032-nucleotide sequence of the constructed whole provirus genome with two LTRs, together with the cellular flanking sequences.

DISCUSSION

Provirus Structure. The LTR structure (U3-R-U5) is thought to play essential roles in integration of provirus DNA into the host chromosomal DNA and also in regulation of transcription of the provirus genome (10, 11). The provirus DNA in λATK-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: ATL, adult T-cell leukemia; ATLV, ATL virus; HTLV, human T-cell leukemia virus; LTR, long terminal repeat.

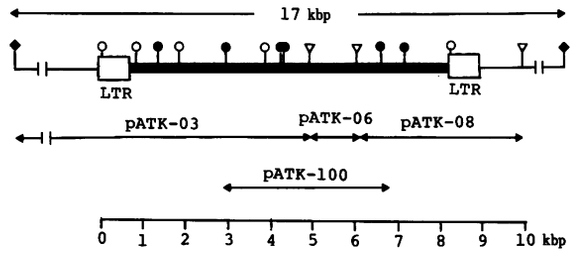

FIG. 1. Restriction map of ATLV provirus clones. The provirus DNA is shown by the thick line with a LTR (box) at each end. The positions of the inserts from clones pATK-03, pATK-06, pATK-08, and pATK-100 are shown under the full provirus genome in λATK-1. ◆, *Eco*RI; ○, *Sma* I; ●, *Pst* I; and ▽, *Bam*HI. kbp, Kilobase pairs.

1 contained two direct repeats of the LTR sequence, one at each end, and the structural features were similar to those in λATM-1, which was isolated from cell line MT-1 (7). Comparison of these two clones revealed the following features. (*i*) Sequences of the LTRs are identical except for 6 base changes at positions 38, C to T; 90, G to A; 146, A to G; 209, G to A; 316, A to G; 481, G to A; and one base (A) insertion at position 190. (*ii*) Cellular flanking sequences are directly repeated by 6 bases in both clones, but the sequences themselves are different, reflecting different integration sites (Fig. 3). Previously, we reported 7-base direct repeats of cellular sequences in λATM-1, but careful reinvestigation demonstrated that there are in fact 6-base repeats. (*iii*) The lengths of the viral sequences between the two LTRs are identical within the limits of experimental errors, although the nucleotide sequence of λATM-1 was not fully determined. The above results indicate that two clones, from cell line and leukemia blood cells, represent a similar ATLV genome.

The unique structures of the LTR previously reported (7) have also been confirmed in this paper. These are (*i*) the extremely long size of R (terminally redundant sequence of genomic RNA) with 229 bases and (*ii*) the absence of the poly(A) signal around the poly(A) site, which is the end of R. With few exceptions, all eukaryotic mRNA containing poly(A) contained the poly(A) signal A-A-T-A-A-A at 10–30 bases upstream of the poly(A) site, but from the sequence of ATLV LTR, we speculated in the previous paper (7) that the poly(A) signal is dispensable for polyadenylylation. However, the nucleotide sequence in the LTR was found to be arranged into a possible secondary structure (Fig. 4), which may explain why transcription terminates within the 3′ LTR but does not terminate in the 5′ LTR. In the 3′ LTR, the RNA transcript that had been initiated at the 5′ LTR would form a hairpin structure, as shown in Fig. 4; thus, the poly(A) signal A-A-T-A-A-A, which is located before the "TATA" box or at 276 bases upstream of the poly(A) site, is arranged into 20 bases before the poly(A) site. In this structure, the signal A-A-T-A-A-A might become effective in the RNA level. But in the 5′ LTR, transcription starts from the cap site, which is located in the loop; therefore, the RNA transcript lacks the poly(A) signal, thus allowing further transcription. A model for inactivation of the A-A-T-A-A-A signal by a possible secondary structure was also proposed in the LTR of murine leukemia virus by Benz *et al.* (12). Our model for ATLV suggests that signals separated by a long nucleotide sequence could be aligned into functional form by conformational rearrangements; therefore, a definite structure in the primary sequences might not necessarily be required. However, this could be an exceptional case.

Capacity of the Genome To Code the Proteins. In general, replication-competent retroviruses have a common gene organization that is *gag*, *pol*, and *env* in this order from the 5′ end of the genomic RNA (13). The DNA sequence of ATLV contained three large open reading frames and four additional smaller ones (Fig. 2). Other possible open frames in the various phases are <200 bases, corresponding to a coding capacity for 70 amino acids. The three large reading frames probably correspond to *gag*, *pol*, and *env* because of their positions and for reasons discussed later.

gag *gene*. The first open frame, which starts from the ATG codon at position 802 and terminates with TAA at position 2,089, could code for a 48,000-dalton protein consisting of 429 amino acids. The recently reported NH_2-terminal sequence of 25 amino acids of p24 in HTLV (14), which is similar to ATLV (5), is identical to a part of this 48,000-dalton protein, which starts from proline at position 1,192, as marked in Fig. 2. The COOH terminus of p24 of HTLV is leucine (14) and this may correspond to the leucine at position 1,831. The predicted p24 of ATLV has a molecular mass of 23,940 daltons and its amino acid composition is very similar to that of p24 of HTLV reported by Oroszlan *et al.* (Table 1) (14). This finding is direct evidence that p24 is virus encoded and also is consistent with the fact that an antibody against p24 of HTLV is crossreactive with ATLV antigens (15). Thus, the first large open frame appears to be the *gag* gene coding for a *gag*-precursor protein, $Pr48^{gag}$. To form p24, the $Pr48^{gag}$ should be cleaved into at least three proteins—that is, a 14,000-dalton protein from the NH_2-terminal, a 24,000-dalton protein from the middle, and a 9,000-dalton protein from the COOH terminal portions of the $Pr48^{gag}$. The molecular masses of the presumed polypeptides may correspond to the 17,000-, 24,000-, and 11,000-dalton proteins, within the limits of experimental errors; these proteins were found previously to be associated with ATLV virions (4).

pol *gene*. In animal retroviruses, the *pol* gene is located after the *gag* gene and is translated into the *gag–pol* polyprotein by changing the reading frame after splicing of the genomic RNA (ref. 16) or by suppressing one termination codon, which appears after the *gag* gene in the frame (17). Because ATLV has the general structural features of the retrovirus genome, such as LTR structure and tRNA binding site (7), it is reasonable to expect that ATLV has the usual gene organization. Thus, the second reading frame from GGC at position 2,498 to TAA at position 5,185 is expected to be the *pol* gene coding for reverse transcriptase. This is the largest open frame and it can code for

Table 1. Amino acid composition of p24

Amino acid	p24 of ATLV	p24 of HTLV*
Asn	9	{21†
Asp	10	
Thr	9	10
Ser	13	14
Gln	21	{36‡
Glu	9	
Pro	18	22
Gly	11	15
Ala	20	24
Cys	3	—
Val	9	7
Met	4	4
Ile	8	8
Leu	28	32
Tyr	5	6
Phe	4	5
His	8	9
Lys	10	12
Arg	11	11
Trp	4	—

* Oroszlan *et al.* (14).
† Asn and Asp.
‡ Gln and Glu.

5'-LTR
gag
pol

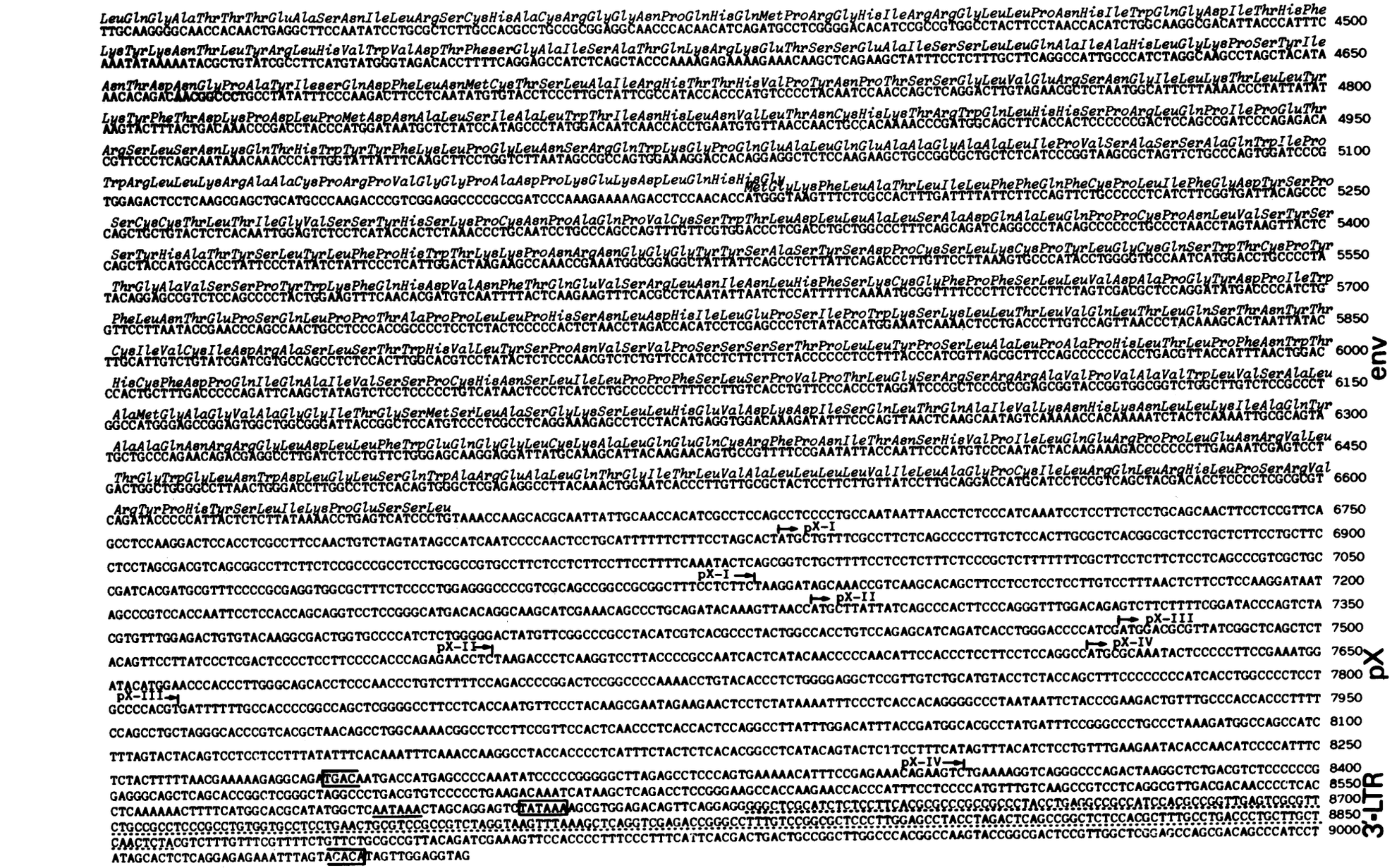

FIG. 2. Complete nucleotide sequence of the ATLV provirus genome. The nucleotides are numbered from the 5′ end of the 5′ LTR. The sequence is translated into amino acids in the region from positions 802 to 2,088 and positions 2,498 to 6,043, and the remaining smaller open frames are marked by small vertical lines. All open frames are started by ATG, except the putative *pol* gene, which started with GGC. [ ], LTR; ▭, TATA box; ——, poly(A) signal; ------, the sequence R; and ■, the NH_2- and COOH-terminal amino acids reported for p24.

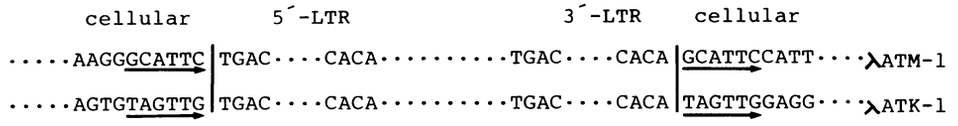

FIG. 3. Nucleotide sequences of the virus–cellular junction in the two clones λATK-1 and λATM-1.

896 amino acids, corresponding to a 99,000-dalton protein. This molecular mass is similar to that of the known reverse transcriptase, but we could not define the NH_2 terminus, because no structural information on the enzyme of ATLV or HTLV is available. Because there are several termination codons in every reading frame after the *gag* gene [at positions 2,089, 2,161, 2,182, 2,239, 2,257, 2,272, 2,347, 2,422, 2,455, and 2,495 in the frame for *gag* and *pol* (frame I), positions 2,123, 2,186, 2,198, 2,288, and 2,438 in frame II, and positions 2,316, 2,370, 2,466, 2,418, and 2,448 in frame III], splicing of the genomic RNA is expected to eliminate the stop codons to read through *gag* to the putative *pol* gene, although we have no evidence for a possible presence of a polyprotein of *gag–pol*.

env *gene*. The third large open frame, which starts at the ATG codon at position 5,180 and terminates with the TAA codon at position 6,644, has the capacity to code for a 54,000-dalton protein composed of 488 amino acids. This frame and the predicted amino acids have the following features in common with the *env* gene products of animal retroviruses. (*i*) The ATG codon at position 5,180 for initiation of the 54,000-dalton protein is located within the putative *pol* gene overlapping by 5 bases. Similar overlappings between *pol* and *env* are also observed in Rous sarcoma virus (D. Schwarz, R. Tizard, and W. Gilbert, personal communication) and murine leukemia virus genomes (18). (*ii*) About 20 amino acids of the NH_2-terminal portion are rich in hydrophobic residues, and this characteristic is similar to that of signal peptides proposed for the *env* gene product of Rous sarcoma virus and murine leukemia virus (18). (*iii*) The 54,000-dalton protein contains five possible sites for glycosylation—that is, Asn-X-Thr/Ser sequences (19) at positions 5,597, 5,843, 5,909, 5,993, and 6,389. Because the *env* gene products are generally glycoproteins, presence of the sites for glycosylation is expected to be essential, although it may not be enough. The product of the *env* of ATLV or HTLV has not been identified, but the characteristics of the putative 54,000-dalton protein described above suggest that this open frame is the *env* gene rather than the *onc* gene.

Other genes? In addition to *gag*, *pol*, and *env*, the ATLV sequence determined has four extra open frames, as indicated in Fig. 2, which have capacities to code for proteins pX-I to pX-IV, with molecular masses of 11,000, 10,000, 12,000, and 27,000 daltons, respectively. Although the presence of these proteins in infected or leukemia cells remains to be studied, some of them might have functions in the process of transformation of infected T cells. If some of these sequences have the common features with the known *onc* genes in acute leukemia viruses, similar nucleotide sequences are expected to be present in normal human DNA. However, the subcloned DNA fragment containing this region did not significantly hybridize with normal human DNA in Southern blotting analysis. This preliminary result indicated that the region containing four extra open frames is not homologous with the human c-*onc* genes. Similar experiments using the other parts of viral DNA fragments suggested that ATLV has no *onc* gene derived from the human genome; however, it is possible that ATLV may contain a gene that is involved in induction of abnormal T-cell proliferation but not derived from the human DNA.

Finally, it should be pointed out that the predicted viral genes or gene products could be tentative, because the provirus analyzed in this paper is that integrated in leukemia cells, and we have no direct evidence for the replicative competence of this provirus, including the viral infection.

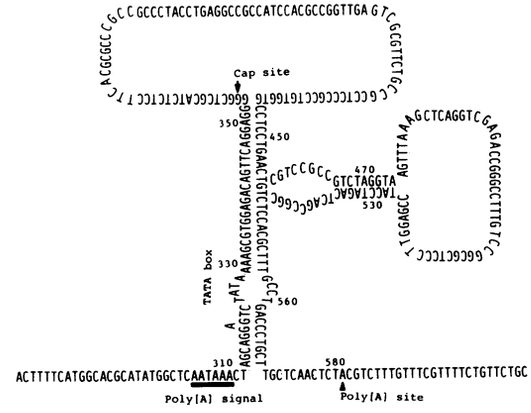

FIG. 4. Possible secondary structure of the nucleotide sequence around the cap site and poly(A) site in the LTR.

The authors thank Dr. H. Sugano for valuable discussion and encouragement during this work. This work was supported in part by a Grant-in-Aid for Cancer Research from the Ministry of Education, Science and Culture of Japan.

1. Poiesz, B. J., Ruscetti, F. W., Gazdar, A. F., Bunn, P. A., Minna, J. D. & Gallo, R. C. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 7415–7419.
2. Reitz, M. S., Poiesz, B. J., Ruscetti, F. W. & Gallo, R. C. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 1887–1891.
3. Hinuma, Y., Nagata, K., Hanaoka, M., Nakai, M., Matsumoto, T., Kinoshita, K., Shirakawa, S. & Miyoshi, I. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 6476–6480.
4. Yoshida, M., Miyoshi, I. & Hinuma, Y. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 2031–2035.
5. Popovic, M., Reitz, M. S., Sarngadharan, M. G., Robert-Guroff, M., Kalyanaraman, V. S., Nakao, Y., Miyoshi, I., Minowada, J., Yoshida, M., Ito, Y. & Gallo, R. C. (1982) *Nature (London)* **300,** 63–66.
6. Uchiyama, T., Yodoi, J., Segawa, K., Takatsuki, K. & Uchino, H. (1977) *Blood* **50,** 481–492.
7. Seiki, M., Hattori, S. & Yoshida, M. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 6899–6902.
8. Blattner, F. R., Blechl, A., Denniston-Thompson, K., Faber, H. E., Richards, J. E., Slightom, J. L., Tucker, P. W. & Smithies, O. (1978) *Science* **202,** 1279–1284.
9. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
10. Hsu, T. W., Sabaran, J. L., Mark, G. E., Guntaka, R. V. & Taylor, J. M. (1978) *J. Virol.* **28,** 810–818.
11. Varmus, H. E. (1982) *Science* **216,** 812–820.
12. Benz, E. W., Jr., Wydro, R. M., Nadal-Ginard, B. & Dina, D. (1981) *Nature (London)* **288,** 665–669.
13. Vogt, P. K. (1977) *Compr. Virol.* **10,** 341–455.
14. Oroszlan, S., Sarngadharan, M. G., Copeland, T. D., Kalyanaraman, V. S., Gilden, R. V. & Gallo, R. C. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 1291–1294.
15. Kalyanaraman, V. S., Sarngadharan, M. G., Nakao, Y., Ito, Y., Aoki, T. & Gallo, R. C. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 1653–1657.
16. Weiss, S. R., Hackett, P. B., Opperman, H., Ullrich, A., Levinthow, L. & Bishop, J. M. (1978) *Cell* **15,** 607–614.
17. Philipson, L., Andersson, P., Olshevsky, U., Weinberg, R. & Baltimore, D. (1978) *Cell* **13,** 189–199.
18. Shinnick, T. M., Lerner, R. A. & Sutcliffe, J. G. (1981) *Nature (London)* **293,** 543–548.
19. Chen, R. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 5788–5792.